

The Enterprise AI Factory

From Ambition to Production with HPE Private Cloud AI

Mladen Vukadinović, Solutions Sales Specialist

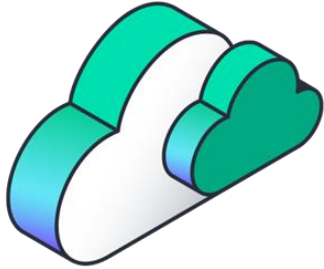
vForum Adria 2026 · Opatija · 29 September 2026



AI is transforming every industry

But a chasm exists
between ambition and
execution

Where do you start?



Public Cloud

Rapid scalability; global reach

Cons:

Data privacy
Performance
Infrastructure control



Private Cloud

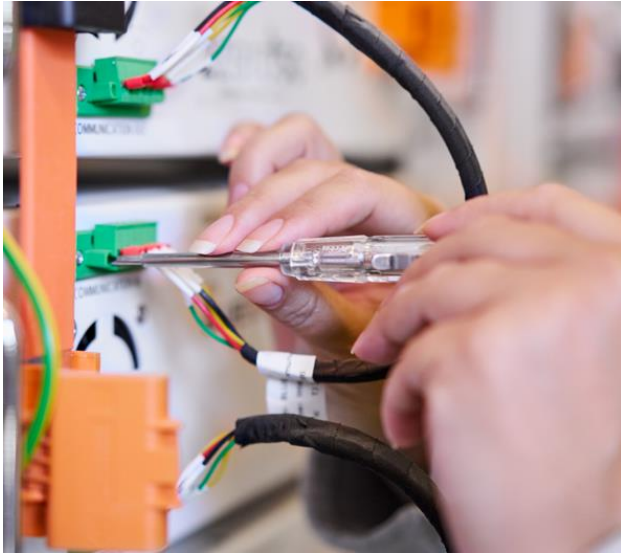
Infrastructure, security & compliance control
Reliable performance

Cons:

Specialized talent
Configuration, management, maintenance by team

*Private cloud's growth driven by need for enhanced security, data sovereignty,
and control to train and deploy AI models using the organization's sensitive and valuable data*

How you deploy private cloud matters



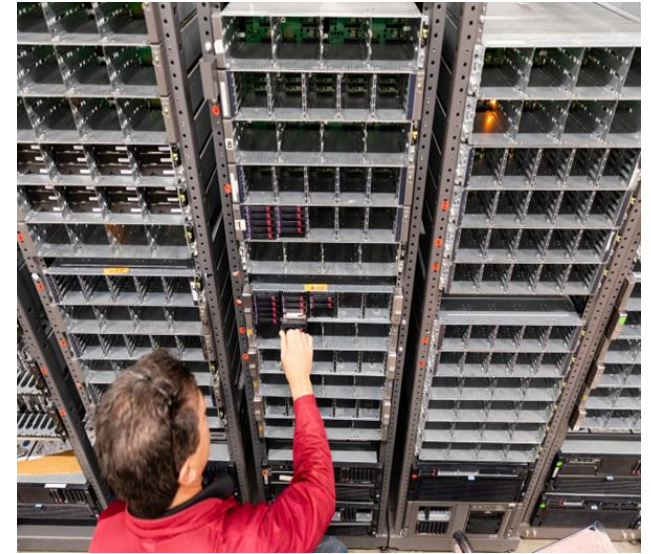
Build-it-yourself

Greatest complexity,
long time-to-value



Reference architecture + services

Moderate complexity,
moderate time-to-value



Turnkey solution

Simplest and fastest
time-to-value

From my own lab: what I built

Planning assistant + RAG

An existing system that plans from database records: parts to spare parts, contracts, break-and-fix delivery.

I added RAG so people can simply ask questions in plain language.

Multi-agent team

Several agents working together, each with its own "good enough" model for its task.

Right-sized models instead of one giant model for everything.

Data pipelines

Daily pipelines that keep the data behind it fresh and consistent.

Most of the effort goes into data, not into the model.

The demo worked in days. Scaling it out with control is a different story.

Demo is easy. **Scaling with control** is the real work.

What I had to solve myself

Securing the platform: physical hosts, iLO, hypervisor or Kubernetes

Who sees which data: data scientists, engineers, users

Giving external partners access while hiding sensitive data

Who can use which models, and which types of models

Changing code and models without disrupting work or crashing infrastructure

Running a right-sized model for each task, without one workload starving the others

Keeping the whole ecosystem patched and supported

How HPE Private Cloud AI handles it

Stack integrated and hardened at the factory, zero-trust security

Role-based access and isolated workspaces from one console

Access defined per project or even per single data source

Permissions per model, resource and data set

Automated lifecycle management and staged roll-outs

Several models side by side, GPU quotas and resource allocation

One integrated stack, updates and support from HPE

Too few specialists: **let people do their own job**

Engineers

Instead of
becoming experts in every component: GPUs,
Kubernetes, storage, networking, AI tooling

They can
**oversee one integrated platform from one
console**

Data people

Instead of
configuring every part and troubleshooting why the
pieces don't work together

They can
**work on the data and models – their actual
job**

What I learned: the most important thing is the data – bringing it together, cleaning it, keeping it fresh.
That's where scarce experts should spend their time – not on plumbing.

What is an AI factory?

Delivers a scalable, repeatable engine for innovation

AI-optimized infrastructure – delivers a manufacturing plant for AI

Transforms raw data into actionable intelligence

Accelerates time-to-value

Manages the entire AI lifecycle

AI projects become repeatable business ops



HPE Private Cloud AI

Engineered AI Factory accelerates AI Initiatives

Co-engineered with NVIDIA



HPE's Turnkey AI Factory

Your fast track to AI – from concept to production



Distinct alternative to public cloud and “build-it-yourself”

Turnkey, engineered, full-stack solution

Private cloud experience with public cloud benefits

Focus on the entire AI lifecycle

Strong ecosystem and partner network

Clear economic and performance benefits

Right-sized for every AI journey

Ready-to-go, but flexible

Power of an integrated stack

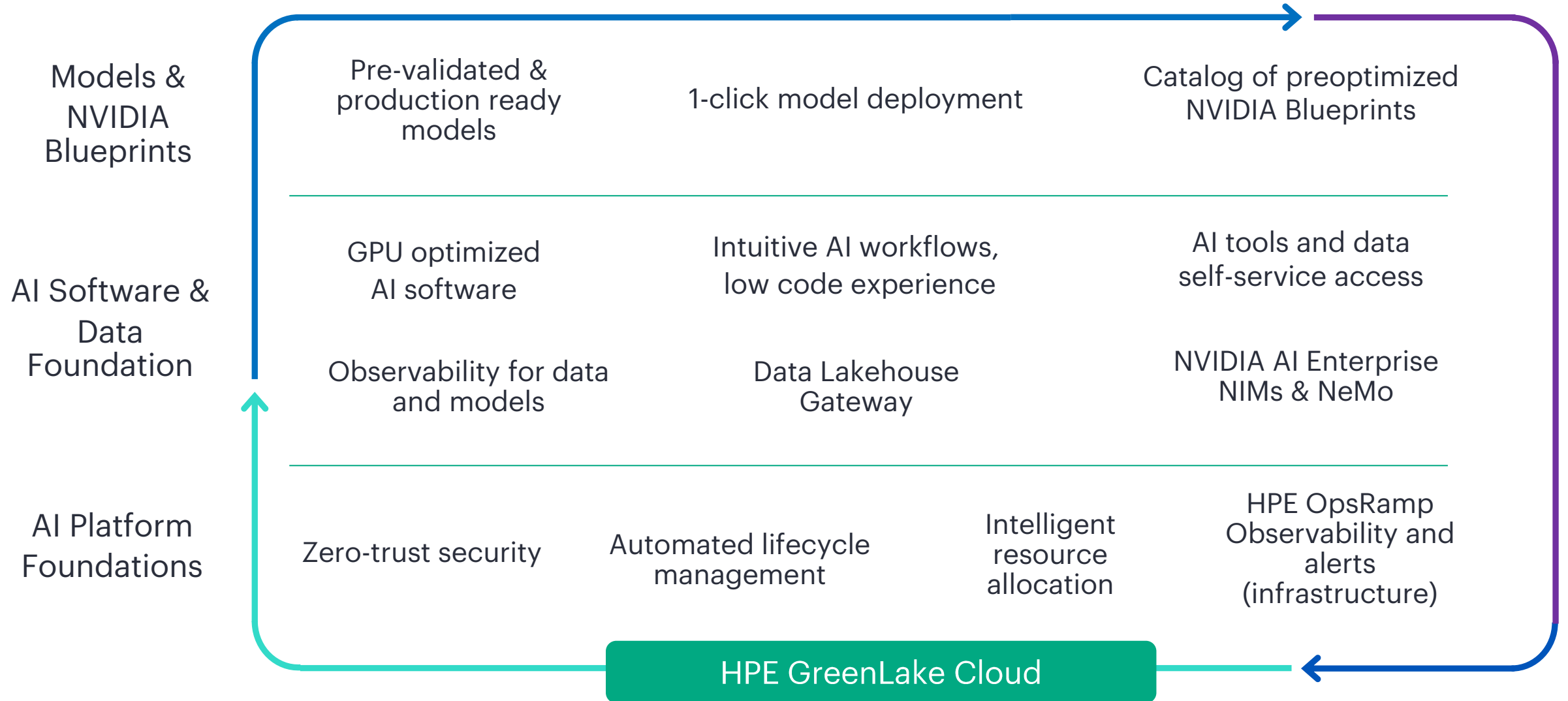
Engineered for performance



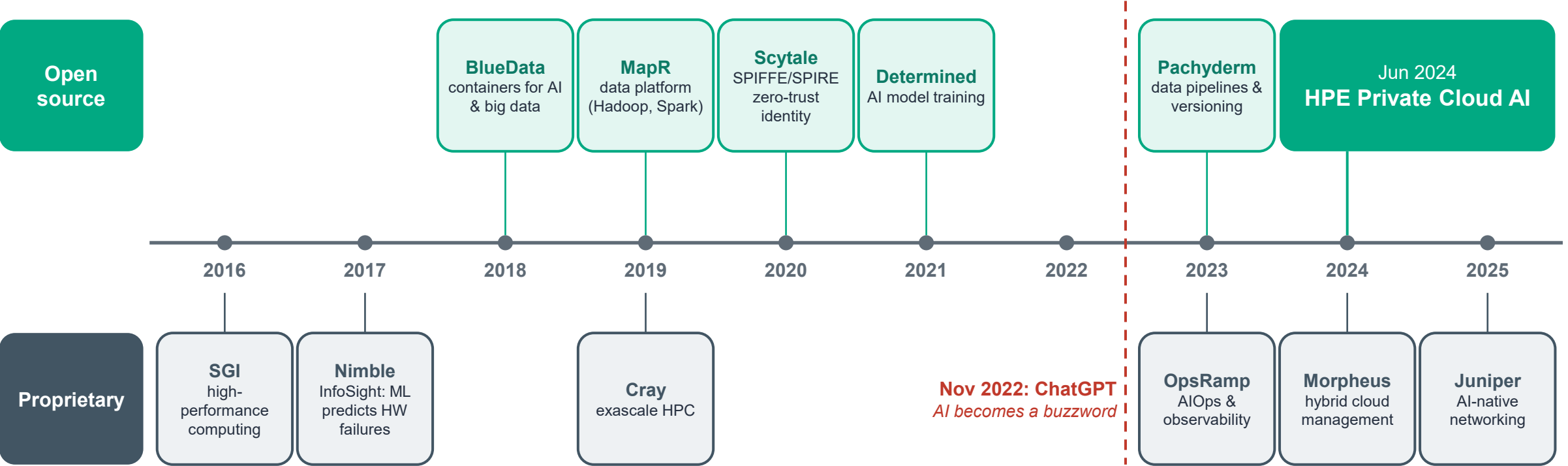
Unified management
experience



Accelerate **your AI journey to production**



Before it was a buzzword: HPE's road to the AI factory



Experience, not a bandwagon: HPE was building this long before the hype.



Open foundations, **enterprise-ready by HPE**

What HPE adds

Integration & lifecycle updates

One console

Access control & multi-tenancy

Zero-trust security

Monitoring

Air-gapped option

Enterprise support

NVIDIA AI Enterprise

NIM microservices

NeMo

Blueprints

Open-source data & ML tools

Kubernetes

Spark

Airflow

Presto

Superset

Kubeflow

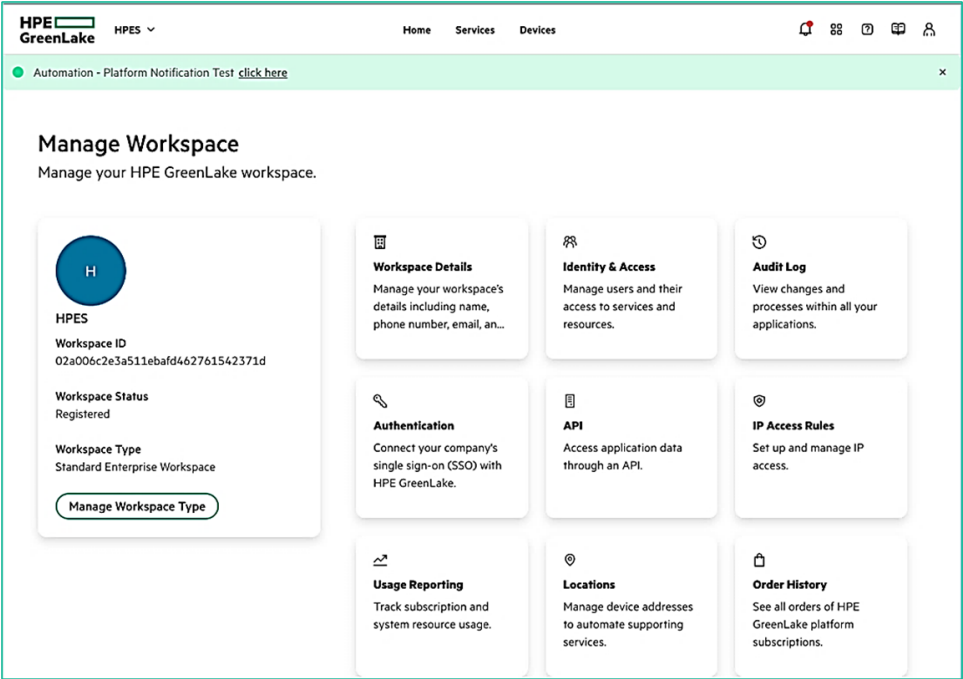
MLflow

Ray

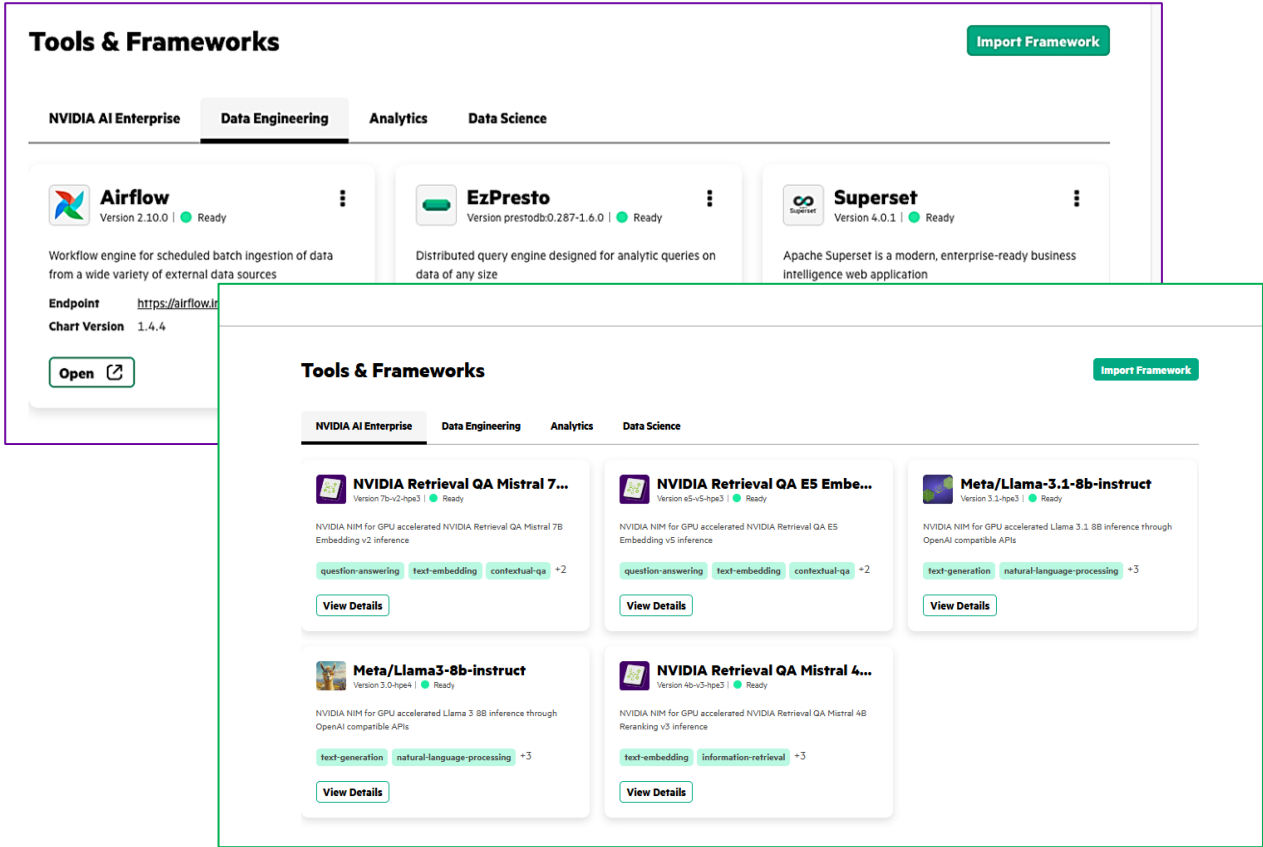
Draws on HPE's AI software heritage: Ezmeral, Determined AI, Pachyderm



Powering innovation across any enterprise



ITOps – unified console simplifies management, security & RBAC



Data science practitioners - Seamless access to tools and models accelerate initiatives



Low-code simplicity

Selected Model*

 Meta/Llama3-8b-instruct
Version 3.0

Select Another

Icon*

 Change Icon

Name*

Virtual Assistant 1728051844847

Description*

A cutting-edge AI solution for creating high-quality, contextually relevant text content suitable for various

Tags

chat × language generation ×

large language models × text-to-text ×

NVIDIA AI Enterprise Supported ×

+ Select Another Data Source

Data Volumes

Source	user
Path	ezmeral_support_doc

Chunk Size

Amount of data processed together; affects speed, memory.

512

Chunk Overlap

Number of data shared between consecutive chunks.

100

GPU*

Number of GPU devices.

1

CPU*

CPU is measured in cores.

12

Memory*

Memory is measured in GiB.

32Gi

Advanced Settings

Maximum Tokens*

2048

Temperature

Frequency

0

Environment Variables

Name

label name

Start canary roll out

Canary roll outs enable you to test your model with a model that is currently deployed. Set your parameters and we'll handle the rest. [Learn more about canary roll outs.](#)

Target Deployment

Which packaged model do you want to serve? ⓘ

 llama-3.1-70b-instruct

2 versions ▾

Which version of this model do you want to serve? ⓘ

 Version 2 (latest) ▾

How much of the traffic should this model see? ⓘ

16%



Quick selects: [5%](#), [25%](#), [50%](#), [75%](#), [100%](#)

Cancel

Done

HP E

Air-gapped option: full AI capability, zero external exposure

Main Theme	
HPE Private Cloud AI – Air-gapped	
High Level Value	Value Prop
Security and Data Sovereignty	Ensures sensitive data remains protected by delivering private AI model customization and inference in air-gapped environments, eliminating exposure to external networks.
Regulatory Compliance	Provides robust security and compliance features, ideal for industries with strict regulatory or sovereignty requirements, such as government, defense, healthcare, finance, and critical infrastructure.
Enhanced AI Functionality	Supports advanced AI development capabilities including inference, fine-tuning, Retrieval Augmented Generation (RAG), and Agentic workflows with full private control of data.
Resilience and Reliability	Operates independently of external connectivity, ensuring uninterrupted service in remote or high-security environments.



The TCO Advantage: HPE Private Cloud AI vs. Public Cloud

Potential savings and key benefits:

- Easier to manage
 - Reduces manual setup and IT resources
 - More efficient infrastructure = reduced TCO
- Faster AI deployment and agility
 - Integrated tools and flexible architecture
 - Quickly adapt to changing market needs
- Enhanced security and compliance
 - Zero-trust security
 - Built-in governance & compliance tools

Validated Benefits of HPE Private Cloud AI



**Up to 82% Lower TCO
and \$8.9M Savings
Compared to Public Cloud
over 3 years**

**Up 50% Lower TCO
and \$3.7M Savings
compared to DIY Solutions
over 3-years**

Source: Enterprise Strategy Group, "[The economic benefits of HPE Private Cloud AI with NVIDIA AI Computing by HPE](#)," March 2025.

The business case depends on useful usage

Demand

How many users, how often, and what response time?

Full operating cost

Infrastructure, software, support, power and the people running it

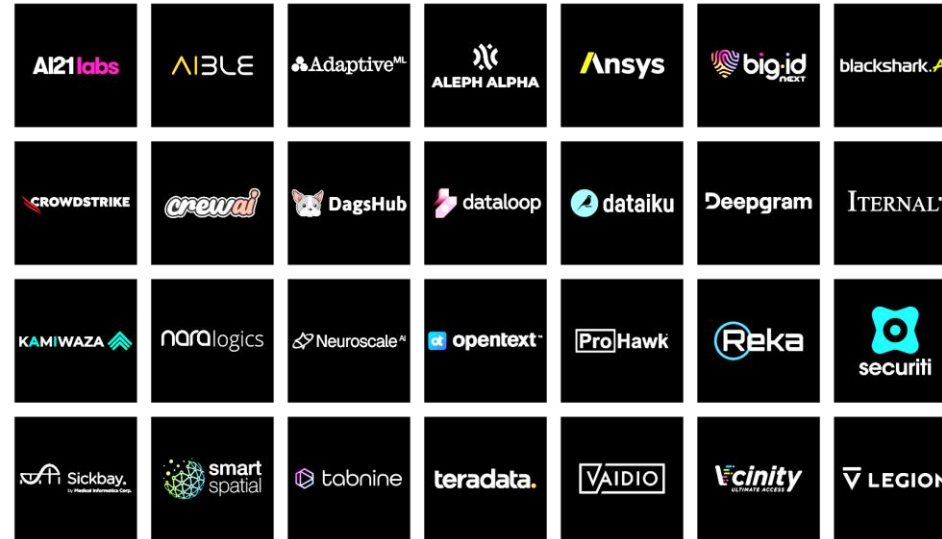
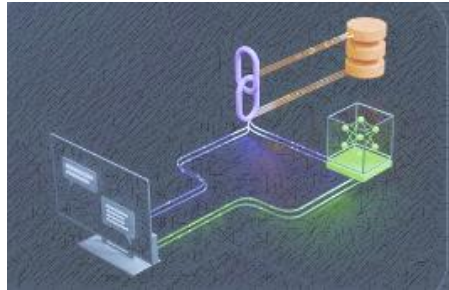
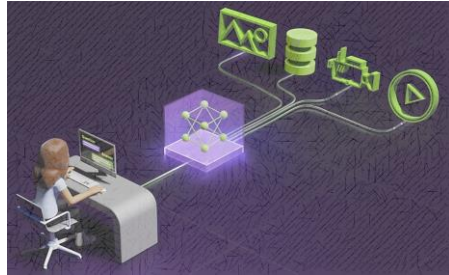
Business result

A useful answer or completed task, measured against the current process

HPE's cited economic studies are scenarios, not a savings guarantee.



Launch pad for your AI ecosystem




accenture
AI Refinery™

ZORA AI™ BY
Deloitte.

NVIDIA

Blueprints

Support for #1
reasoner AIQ



Unleash AI
ISV Partner Program
More than 75 validated AI use
cases

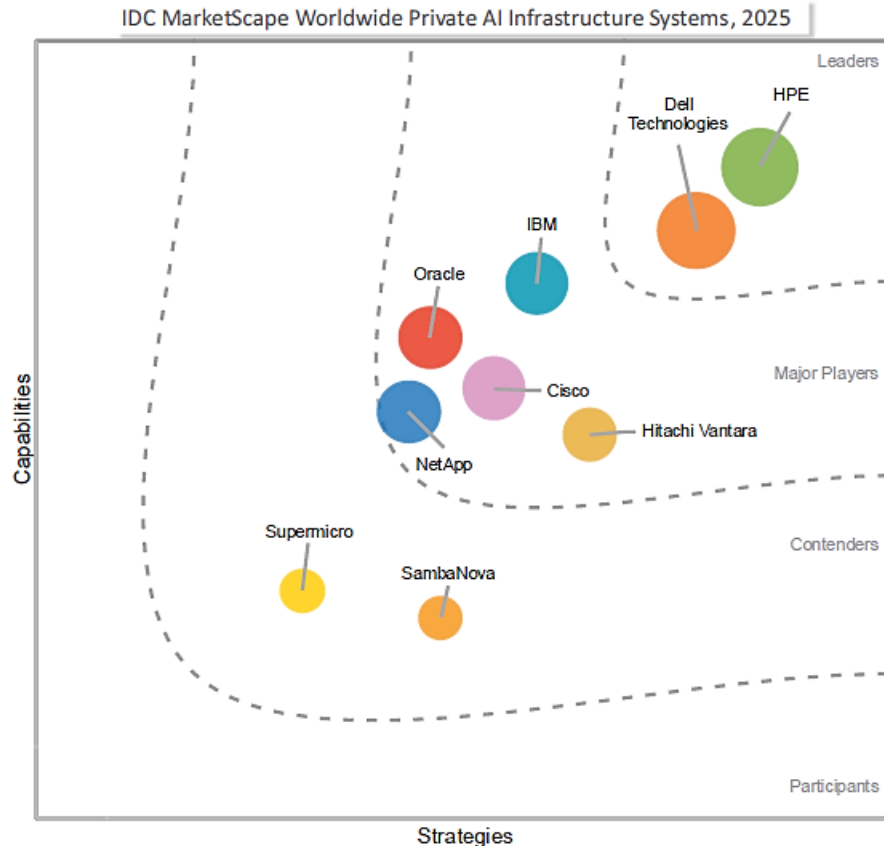


GSI Partnerships
in Action



IDC MarketScape for AI Infrastructure Systems

IDC MarketScape Worldwide Private AI Infrastructure Systems Vendor Assessment



HPE is positioned as a Leader in this 2025 IDC MarketScape for worldwide private AI infrastructure systems

“For customers that want a tightly integrated stack that includes robust storage and management integrations, the HPE Private Cloud AI reduces the complexity related to ensuring data pipelines and AI workloads are efficiently integrated.”

IDC MarketScape vendor analysis model is designed to provide an overview of the competitive fitness of technology and suppliers in a given market. The research methodology utilizes a rigorous scoring methodology based on both qualitative and quantitative criteria that results in a single graphical illustration of each supplier's position within a given market. The Capabilities score measures supplier product, go-to-market and business execution in the short-term. The Strategy score measures alignment of supplier strategies with customer requirements in a 3-5-year timeframe. Supplier market share is represented by the size of the icons.

A practical starting point

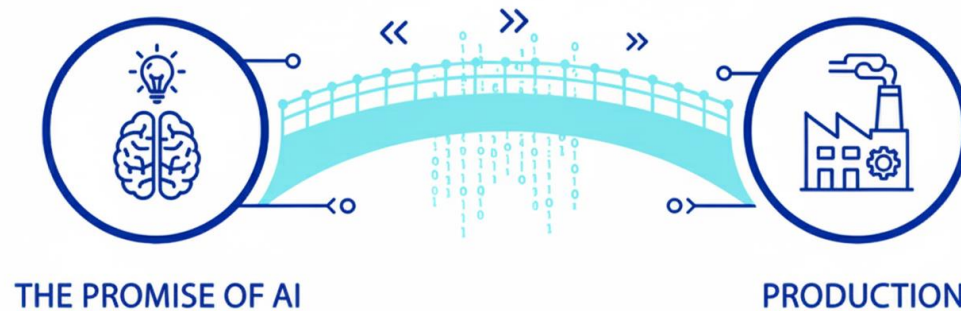
One use case. Real data. A named owner.

Test the answer quality and access rules.
Measure demand and operating effort.
Agree what “ready for production” means.

Bring a use case to an HPE / Selectium workshop or demo.



Key takeaways



- ✓ Fast Time-to-Value: Go from idea to production in weeks, not months.
- ✓ Total Control: Complete security, data sovereignty, and compliance.
 - ✓ Predictable Costs: No more surprise cloud bills.
- ✓ Enterprise Ready: A full-stack, scalable solution, from infrastructure to models.
- ✓ Next step: pick one use case – let's run a workshop or live demo together.

Thank You

Questions?

Mladen Vukadinović (HPE, operated by Selectium)
m.vukadinovic@selectium.com

